DOCUMENT RESUME

ED 220 517 TM 820 534

| | |
|---|---|
| AUTHOR | Cummings, Corenna C. |
| TITLE | Estimates of Multiple Correlation Coefficient Shrinkage. |
| PUB DATE | Mar 82 |
| NOTE | 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982). |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Correlation; Error of Measurement; *Mathematical Formulas; *Multiple Regression Analysis; *Predictive Measurement; Predictor Variables; Research Methodology; Sampling; *Validity |
| IDENTIFIERS | Criterion Variables; *Cross Validation; Sample Size |

ABSTRACT

The accuracy and variability of 4 cross-validation procedures and 18 formulas were compared concerning their ability to estimate the population multiple correlation and the validity of the sample regression equation in the population. The investigation included two types of regression, multiple and stepwise; three sample sizes, N = 30, 60, 120; and, two sets of predictors, n = 4, 8. The results indicate that most of the formulas and cross-validation procedures advocated in the literature were inaccurate and extremely variable. However, dependent upon the type of regression and the population value estimated, conservatively accurate estimators with low variability were identified. (Author/CM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED220517

# ESTIMATES OF MULTIPLE CORRELATION COEFFICIENT SHRINKAGE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. C. Cummings

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

By

Corenna C. Cummings, Ph.D.
Rush University

Paper presented at the Annual Meeting of the
American Educational Research Association
March 19-23, 1982, New York

TM 820 534

# ESTIMATES OF MULTIPLE CORRELATION COEFFICIENT SHRINKAGE

Corenna C. Cummings, Ph.D.
Rush University

The accuracy and variability of four cross-validation procedures and eighteen formulas were compared concerning their ability to estimate the population multiple correlation and the validity of the sample regression equation in the population. The investigation included two types of regression, multiple and stepwise; three sample sizes, N = 30, 60, 120; and, two sets of predictors, n = 4, 8. The results indicate that most of the forumulas and cross-validation procedures advocated in the literature were inaccurate and extremely variable. However, dependent upon the type of regression and the population value estimated, conservatively accurate estimators with low variability were identified.

Cross-validation is used to determine the stability of a sample regression equation in future samples. It has traditionally been a recommended procedure because regression techniques applied to random independent and dependent variables result in the inclusion of chance errors in the sample regression equation. This inclusion of errors results in a correlation coefficient that is an over-estimate of the true relationship in the population. Cross-validation requires that a portion of the sample data be withheld from that data used to calculate the regression equation. Subsequently, the regression equation is applied to the withheld data to obtain the cross-validity coefficient. The regression weights are less stable than they would be if all the sample data were used in the regression procedure since the stability of the regression weights increases as sample size increases.

Alternatives to cross-validation which would not only allow the use of all the data, but would also provide accurate estimates of the stability of the sample regression equation in future samples have been proposed. The proposed alternatives are formulas which use the values of the sample size, the number of independent variables, and the sample multiple correlation to estimate the desired population values. The many formulas which have been proposed have

been designed to estimate either $\rho_c^2$, the squared validity of the sample regression equation in the population, or $\rho^2$, the squared population multiple correlation. A useful theoretical solution to the estimation of these population values has not been found and the available empirical studies are more a source of confusion than of clarification.

This study was designed to compare the proposed formulas and cross-validation techniques in terms of their ability to provide accurate and useful estimates of $\rho_c^2$ and $\rho^2$. In addition, the effects of sample size, type of regression, and number of predictors on the accuracy and variability of the estimates were considered. The results of this study are useful primarily to practical researchers who are using regression procedures with relatively small sample sizes. The results can aid researchers in determining whether to conduct conduct cross-validation procedures or to use a formula estimate to determine the population multiple correlation and the stability of the regression equation in future samples. The accuracy and stability of prediction procedures is extremely important in the field of education most notably in decision-making situations such as selection, placement, and program planning and evaluation.

## Theoretical Background

Concerning human behavior, it is commonly believed that the best predictor of future behavior is past behavior. In education and many other fields, we attempt to quantify this relationship by calculating a regression equation. A sample for which measurements on several independent predictor variables (Xs) and a dependent criterion variable (Y) are known is drawn from an identifiable population, then regression techniques are employed to quantify the relationship between the X variables and the Y variable. It is assumed that in future samples from the same population, the measurements on the independent variables will be known and the goal is to predict the measurement on the dependent variable which

is unknown.

One method for deriving the relationship between the X and Y variables is least squares estimation, the basis of which is to minimize the sum of squared differences between the obtained values of the dependent variable and the predicted values of the dependent variable. The least squares process results in the selection of weights that maximize the multiple correlation between the linear combination of the X variables and the Y variable.

Least squares multiple regression procedures include in the calculated regression equation all the n variables the researcher enters in the calculation of the problem. The assumption is made that the researcher considers the variables of importance in the prediction. Often, however, researchers do not know which of the available X variables are most important, and therefore, they use a regression procedure that systematically includes and/or excludes variables in the regression equation. One such statistical procedure is known as stepwise regression which systematically selects one variable at a time for inclusion in the regression equation, while at the same time, variables previously selected are examined to determine if they should be retained.

Often, in education, the variables of interest are both random variables as when a final grade, Y, is to be predicted from entrance requirement test scores, Xs. "Overfitting" of the regression equation, due to the fitting of error specific to the sample, results in an over-estimation of the population parameters when the sample multiple correlation is used as the estimate. To accurately use the regression equation, it has traditionally been recommended that a cross-validation study be conducted to determine how much overfitting has occurred. The impact on the overestimation of the population multiple correlation is greater when small sample sizes are used for study. Yet, the very process of cross-validation decreases the sample size further and exacerbates the problem. Thus, in practical settings when sample sizes

are often small, the formula estimates could be extremely advantageous.

## Cross-validation

Once a regression equation relating the independent variables and the dependent variable has been calculated from the original sample, cross-validation techniques enable the researcher to assess how well that equation will function in a second sample drawn from the same population.

In the most common cross-validation pattern, the stability of the regression equation calculated in Sample 1 from Population A is examined as applied to Sample 2 from Population A. Mosier (1951) proposed the double cross-validation pattern wherein the stability of the regression equation calculated in Sample 1 of Population A is examined as applied to Sample 2 from Population A and the stability of the regression equation independently calculated in Sample 2 from Population A is examined as applied to Sample 1 from Population A. If the stability is subjectively judged adequate, the samples are combined, a new regression equation is calculated, and the stability is considered to be the average stability of the two original equations.

Mosier (1951) developed the double cross-validation pattern to address the major problem associated with cross-validation. The problem is that a researcher with limited data wishes first, to obtain the regression equation with as stable weights as possible and second, to have the most accurate estimate of that equation's stability in other samples. If all of the limited data is used to calculate the regression equation, no data will be available on which to determine the stability of the equation. However, if some of the limited data is reserved to determine the stability of the equation, the sample size used to calculate the equation is reduced and thus the regression weights are less stable.

The measure or statistic used to determine the stability of the regression equation is the correlation in the cross-validation sample between the predicted

value of the dependent variable based on the original sample regression equation and the obtained value of the dependent variable. This statistic is denoted $R_c$ and is known as the cross-validity or the validity of the original sample regression equation in another sample. It has long been recognized and makes intuitive sense that $E(R_c)$ will be less than $E(R)$, the expected value of the sample multiple correlation or the expected value of the validity of the original sample regression equation in the original sample. Since the goal of least squares regression is to maximize the correlation between the X variables and the Y variable, the least squares regression solution uses everything that increases the correlation, including errors. Thus, the regression equation is said to be overfitted in that chance errors have been included. When this regression equation is applied to another sample, the expected value of the correlation between the predicted and obtained dependent variables will be less than $E(R)$.

R and $R_c$ are sample values and often we wish to estimate the corresponding population values since we can rarely know them directly. R is a positively biased estimator of $\rho$, the population multiple correlation or the validity of the population regression equation in the population, for much the same reason that it is a positively biased estimator of $R_c$. That is, not only has the true regression equation been fitted, but also chance errors have been included. An accurate estimate of $\rho$ is of interest to the practical researcher engaged in a study which is primarily correlational in nature and predictions based on the calculated regression equation applied to new data is not of interest. The researcher engaged in a study in which predictions based on the calculated regression equation applied to new data is of interest will find $R_c$ and $\rho_c$ more useful. The value of $R_c$ will vary from one cross-validation sample to another, however, the expected value of $R_c$, or the average over many samples, will be approximately equal to $\rho_c$, the validity of the sample regression equation in the population. Since the population regression equation in the population

will function better than the sample regression equation in the population, the value of $\rho$ is greater than the value of $\rho_c$. As in Herzberg (1969), the above relationship may be expressed as

$$E(R_c) \simeq \rho_c < \rho < E(R).$$

The technique of cross-validation was designed to obtain the value of $R_c$, which is a more accurate indicator than R, of the expected stability of the sample regression equation when applied to future samples. However, researchers were reluctant to save a portion of a limited sample for cross-validation. The search for a solution to this problem began with an attempt to estimate how much smaller R would become (how much R would "shrink") or the most probable value of the "shrunken R" when the sample regression equation was applied to a new sample from the same population. Formulas were developed to estimate $\rho^2$, the squared population multiple correlation, which was known to be less than the expected value of $R^2$. For a long time, a distinction was not made between $\rho^2$ estimated by $R^2$ with the use of a shrinkage formula and $\rho_c^2$ estimated by a cross-validation procedure except that they were differently obtained estimates that were basically the same. This misunderstanding continued in common textbooks for nearly twenty years after Nicholson (1948) and Lord (1950) independently noted the differences. Thus, formulas designed to estimate $\rho^2$ were found to overestimate $\rho_c^2$, as indeed they should. Formulas were then developed to estimate $\rho_c^2$ which had been the initial goal of cross-validation.

What has ensued in the literature has been a general state of confusion. In both practical and theoretical studies, it is often difficult or impossible to determine what value the researcher wished to estimate and which of the many formulas or algebraic equivalents has been used. In practical studies, the results are difficult to evaluate if the researcher simply reports a value for "shrunken R." In simulation studies designed to examine the effects of certain conditions on selected estimates of shrinkage, formulas are often misnamed and some of the formulas are evaluated under conditions which are inappropriate for

them. This has resulted in a lack of continuity among the simulation studies and has caused evaluative comparisons among the formulas to be a difficult procedure. For the practical researcher designing a study, there are many obstacles to overcome. Most general measurement and statistics books provide little information, fail to distinguish the correlation and regression models, and do not explain the various shrinkage formulas. Should the researcher consult a single simulation study, there may still be difficulty selecting an appropriate formula and citing it accurately. (For a review of the literature on both the proposed formulas and the empirical studies of the formulas, see Cummings (1980).)

The two elements of interest are first, to determine which formulas provide the most accurate estimate of the population multiple correlation, and second, to determine which formulas may be used as an alternative to cross-validation in the estimation of the validity of the sample regression equation in the population. Further, it is important to determine under what research conditions and under what violations of model assumptions would these findings remain applicable.

## Research Design

The problems associated with regression procedures are greatest when small sizes are used, therefore, the research effort was concentrated on sample sizes of N = 30, 60, and 120 which are representative of many applied studies using regression analysis. Further, previous empirical studies have indicated that the discrepancies between population values and formula estimates tend to increase as the number of predictors increase, though this pattern was not always consistent. In this research, two sets of predictor variables were employed. One set included four subtest scores from the ACT Assessment; the other set included four self-reported high school grades in addition to the ACT scores.

The study was conducted by drawing multiple random samples from populations of real data which were of a nature typically submitted to regression and cross-validation procedures. The scope of the study included two populations; two types of regression, multiple and stepwise; the two sets of predictors, n = 4 and 8; the three sample sizes; and ten replications. The data bases consisted of ACT subtest scores, self-reported high school grades, and first term grade-point averages for two complete classes of entering freshmen at a large university.

Two packaged computer programs, BMDP (PIR and P2R) and SPSS (PEARSON CORR) were used for this study. For each population and set of independent variables, BMDPIR was used to calculate the population multiple regression equation and correlation, $\rho$. For each population with eight independent variables available for inclusion, BMDP2R was used to calculate the population stepwise regression equation and correlation, $\rho$. Random samples were selected using a program which incorporated GGUBS, a basic uniform (0,1) pseudo-random number generator, from the packaged computer program IMSL. The regression weights calculated for each sample were applied to the population from which the sample was drawn and the multiple correlation calculated. This yielded $\rho_c$, the population values of the validity of the sample equation in the population.

Although cross-validation procedures require some data be withheld from that data used to calculate the regression equation, the amount held in reserve is not strictly defined. A research practioner would use as much data as possible to calculate the regression equation while holding a sufficient amount of data in reserve to obtain an accurate, reasonable cross-validity coefficient. Typically, the reserved data is one-third to one-half the total sample.

Few previous empirical research studies have included cross-validation indices and most have been incorrect in their comparisons of formula estimates to cross-validation values. Only one of the previous empirical research

studies (Newman, et al., 1979) correctly compared formula estimates to traditional cross-validation indices using a one-half sample held in reserve. Both Uhl and Eisenberg (1970) and Schmitt et al. (1977) used a second sample for cross-validation which was of the same size as the sample used to calculate the regression equation and provide an $R^2$ for the formula estimates. This is not representative of the situation encountered by the research practitioner. The practitioner has a sample size of N. The choice to be made is between either using a portion of the sample to calculate the regression equation and using the reserved portion for cross-validation or using the total sample to calculate the regression equation and using a formula to estimate the validity of the sample regression equation in the population. Claudy (1978) did not comparatively evaluate traditional cross-validation but developed a new cross-validation procedure which used two, one-half sample cross-validity indices. In view of the various cross-validation procedures available, this study included four cross-validation approaches based on both one-half of the sample and one-third of the sample held in reserve, Mosier's (1951) double cross-validation and Claudy's (1978) newly formulated cross-validation. More specifically, the four cross-validation indices included the following:

1. Half-sample cross validation. The regression equation calculated in one-half of the sample is applied to the other half of the sample and the correlation between the predicted and obtained dependent variables is calculated. This is the traditional procedure for withholding half of the sample for cross-validation. The result is $R_{c\frac{1}{2}}$, the cross-validity index, and may be considered an estimate of $\rho_c$.
2. One-third sample cross validation. The regression equation calculated in two-thirds of the sample is applied to the other third and the correlation between predicted and obtained dependent variables is calculated. This is the traditional procedure for withholding

a third of the sample for cross-validation. The result is $R_{c1/3}$ the cross-validity index, and may be considered an estimate of $\rho_c$.

3. Double cross-validation (Mosier, 1951). The regression equation calculated in one-half of the sample is applied to the second half of the sample and vice versa. Two cross-validity indices are calculated and averaged. The average is $R_{cm}$ and may be considered an estimate of $\rho_c$. The formula is

$$\hat{\rho}_c = R_{cm} = (R_{c1} + R_{c2})/2$$

4. Double cross-validation (Claudy, 1978. The regression equation calculated in one-half of the sample is applied to the second half of the sample and vice versa. Two cross-validity indices are calculated. The two sample multiple correlations and the two cross-validity indices are averaged. The average is $R_{cc}$ and may be considered an estimate of $\rho$. The formula is

$$\hat{\rho} = R_{cc} = (R_1 + R_2 + R_{c1} + R_{c2})/4$$

In previous empirical research studies in this area, the formulas chosen for investigation varied. A formula which appeared to be most accurate in one study appeared less promising in another study when compared to a different set of formulas. No single study has provided an evaluation of all formulas simultaneously. Therefore, this study was designed to compare eighteen formulas and formula combinations.

In the list of formulas given in Appendix A, the number in parentheses to the right of each formula correspond to the designation given in Cummings (1980) and to tables which appear later. The literature citations to the right of the formulas are discussed at length in Cummings (1980).

Following the calculations of the sample and population correlations, the cross-validation indices, and the formula estimates, the following values were calculated: the mean, standard deviation, sum of scores, and sum of

of squared scores for 1) the formula estimates, 2) the formula estimates subtracted from $R^2$, $\rho^2$, and $\rho_c^2$, and 3) the cross-validation indices subtracted from $R^2$, $\rho^2$, and $\rho_c^2$.

## Results and Discussion

The population values of the multiple correlation in the two populations were: for four predictors with multiple regression, .3500 and .3494; for multiple regression with eight predictors, .4774 and .4818; and, for stepwise regression with six variables selected, .4769 and .4809. The predictor variable intercorrelations were in the range of .118 to .687. The range of the sample multiple correlations was .2474 to .7741.

The squared sample multiple correlation would be a very inaccurate estimate of either the squared population multiple correlation or the squared validity of the sample regression equation in the population. The use of any estimate, cross-validation or formula, would generally be more accurage especially when sample size is small even up to N = 120 and the number of predictors is large.

The basic function of the formulas is to reduce $R^2$. The formulas reduce $R^2$ more as the number of predictors increases and less as the sample size increases. There are two exceptions to this pattern which are first, (6) which equals $R^2$ when n = 1; and second, (20) which is greater than $R^2$ when n = 1. The use of formulas to estimate population values is simple compared to the application of cross-validation procedures.

Of the cross-validation procedures, $R_{cm}^2$ and $R_{cc}^2$ requires more work than $R_{c\frac{1}{2}}^2$ or $R_{c\frac{1}{3}}^2$ to obtain since they both use two half-sample cross-validity indices. However, if two half-sample cross validities were calculated to obtain $R_{cm}^2$ as an estimate of $\rho_c^2$, $R_{cc}^2$ could also be easily calculated to estimate $\rho^2$. If $R_{c\frac{1}{2}}^2$ or $R_{c\frac{1}{3}}^2$ were used, there would be one estimate available and it is considered an estimate of $\rho_c^2$. It was shown in the results that reserving half of the sample for cross-validation resulted in more accurate estimates

than did reserving one-third of the sample.

Cross-validation and formula estimates of $\rho_c^2$ were compared for both multiple and stepwise regression to determine their relative accuracy and usefulness. Note that the values of $\rho_c^2$ were obtained by applying the regression weights calculated in the samples to their respective populations. Table 1 shows the mean differences of $\rho_c^2$ minus the estimates for the 1978 and 1979 combined data for both multiple and stepwise regression. Each tabled value for multiple regression is the mean of approximately 120 differences and for stepwise regression is the mean of approximately 58 differences. Table 3 shows the standard deviations of the estimates for multiple and stepwise regression. For multiple regression, formula (7 &23) was the most accurate and least variable of the formula estimates. Having the smallest standard deviation suggests that it has less chance than the other formulas of resulting in a large error and is therefore more useful. Formula (7 & 23) had a slight tendency to overestimate $\rho_c^2$ and this may be considered a disadvantage. When practical researchers need an estimate of the functioning of the regression equation in future samples, they would most often prefer a conservative estimate, that is, one which tends to underestimate $\rho_c^2$ . The most accurate, conservative choice of the formulas available is (15) which unfortunately tended to underestimate $\rho_c^2$ to a great degree and was quite variable. Cross-validation estimates $R_{c\ddagger}^2$ and $R_{cm}^2$ were similar in accuracy to (7 & 23) and tended to underestimate $\rho_c^2$, but were more variable than (7 & 23). If the effort and probable increased cost involved in obtaining cross-validation estimates were not an issue, they would be reasonable conservative choices except for the variability. Therefore, the less conservative choice, (7 & 23), would be best.

For estimates of $\rho_c^2$ with stepwise regression, formula (18) was most accurate of the formulas. It, however, was rather variable and tended to overestimate $\rho_c^2$ . Considerable accuracy would be sacrificed in order to select another formula with less variability. If cross-validation were a

TABLE 1

Mean Differences and Ranks of $\rho_c^2$ Minus Estimates for Multiple and Stepwise Regression

| | Multiple Regression | | Stepwise Regression | |
|---|---|---|---|---|
| | Mean Diff | Rank | Mean Diff | Rank |
| $\rho_c^2 - \hat{\rho}^2$ | | | | |
| $\rho c^2 - Rcc^2$ | -.079 | | -.039 | |
| $\rho c^2 - (5)$ | -.053 | | -.117 | |
| $\rho c^2 - (6)$ | -.071 | | -.131 | |
| $\rho c^2 - (7)$ | -.052 | | -.117 | |
| $\rho c^2 - (9)$ | -.060 | | -.126 | |
| $\rho c^2 - (13)$ | -.057 | | -.124 | |
| $\rho c^2 - (20)$ | -.077 | | -.141 | |
| $\rho c^2 - (21)$ | -.072 | | -.119 | |
| $\rho_c^2 - \hat{\rho}_c^2$ | | | | |
| $\rho c^2 - Rc\frac{1}{3}^2$ | -.033 | 4 * | -.048 | 4 |
| $\rho c^2 - Rc\frac{1}{2}(1)^2$ | .000 | 1 * | .026 | 3 |
| $\rho c^2 - Rc\frac{1}{2}(2)^2$ | .003 | 2 * | -.017 | 1 |
| $\rho c^2 - Rcm^2$ | .010 | 3 * | .020 | 2 |
| $\rho c^2 - (15)$ | .060 | 9 | -.082 | 3 |
| $\rho c^2 - (17)$ | .091 | 11 | -.079 | 2 |
| $\rho c^2 - (18)$ | .080 | 10 | -.066 | 1 |
| $\rho c^2 - (19)$ | -.011 | 4.5 | -.098 | 6 |
| $\rho c^2 - (5\&23)$ | -.004 | 2 | -.097 | 5 |
| $\rho c^2 - (6\&23)$ | -.021 | 6.5 | -.124 | 10 |
| $\rho c^2 - (7\&23)$ | -.002 | 1 | -.096 | 4 |
| $\rho c^2 - (9\&23)$ | -.011 | 4.5 | -.114 | 9 |
| $\rho c^2 - (13\&23)$ | -.010 | 3 | -.110 | 8 |
| $\rho c^2 - (20\&23)$ | -.030 | 8 | -.143 | 11 |
| $\rho c^2 - (21\&23)$ | -.021 | 6.5 | -.101 | 7 |
| $\rho c^2 - R^2$ | -.150 | | -.139 | |

*Cross-validation procedures are ranked separately for each of the situations.

TABLE 2

Mean Differences and Ranks of $\rho^2$ Minus Estimates for Multiple and Stepwise Regression

| | Multiple Regression | | Stepwise Regression | |
|---|---|---|---|---|
| | Mean Diff | Rank | Mean Diff | Rank |
| $\rho^2 - \hat{\rho}^2$ | | | | |
| $\rho^2 - Rcc^2$ | -.013 | | .061 | |
| $\rho^2 - (5)$ | .009 | 5 | -.018 | 1.5 |
| $\rho^2 - (6)$ | -.007 | 2 | -.032 | 6 |
| $\rho^2 - (7)$ | .012 | 6 | -.018 | 1.5 |
| $\rho^2 - (9)$ | .005 | 1 | -.026 | 5 |
| $\rho^2 - (13)$ | .008 | 3.5 | -.025 | 4 |
| $\rho^2 - (20)$ | -.013 | 7 | -.041 | 7 |
| $\rho^2 - (21)$ | -.008 | 3.5 | -.020 | 3 |
| $\rho^2 - \hat{\rho}_c^2$ | | | | |
| $\rho^2 - Rc\frac{1}{3}^2$ | .032 | | .035 | |
| $\rho^2 - Rc\frac{1}{3}(1)^2$ | .065 | | .125 | |
| $\rho^2 - Rc\frac{1}{3}(2)^2$ | .069 | | .083 | |
| $\rho^2 - Rcm^2$ | .074 | | .120 | |
| $\rho^2 - (15)$ | .125 | | .018 | |
| $\rho^2 - (17)$ | .156 | | .020 | |
| $\rho^2 - (18)$ | .144 | | .033 | |
| $\rho^2 - (19)$ | .054 | | .002 | |
| $\rho^2 - (5\&23)$ | .061 | | .002 | |
| $\rho^2 - (6\&23)$ | .043 | | -.025 | |
| $\rho^2 - (7\&23)$ | .062 | | .003 | |
| $\rho^2 - (9\&23)$ | .054 | | -.014 | |
| $\rho^2 - (13\&23)$ | .055 | | -.011 | |
| $\rho^2 - (20\&23)$ | .034 | | -.044 | |
| $\rho^2 - (21\&23)$ | .043 | | -.002 | |
| $\rho^2 - R^2$ | -.086 | | -.039 | |

TABLE 3

Standard Deviations and Ranks of the Estimates
for Multiple and Stepwise Regression

| | Multiple Regression | | Stepwise Regression | |
|---|---|---|---|---|
| | Mean Diff | Rank | Mean Diff | Rank |
| $\hat{\rho}^2$ | | | | |
| $R_{cc}^2$ | .1070 | 2 * | .1264 | 8 |
| (5) | .1168 | 4.5 * | .1062 | 4 |
| (6) | .1142 | 3 * | .1059 | 2 |
| (7) | .1065 | 1 * | .1060 | 3 |
| (9) | .1195 | 7 * | .1091 | 5.5 |
| (13) | .1204 | 8 * | .1091 | 5.5 |
| (20) | .1168 | 4.5 * | .1092 | 7 |
| (21) | .1192 | 6 * | .1056 | 1 |
| $\hat{\rho}_c^2$ | | | | |
| $R_{c\frac{1}{3}}^2$ | .1345 | 13 | .1544 | 15 |
| $R_{c\frac{1}{3}(1)}^2$ | .1154 | 12 | .1080 | 9 |
| $R_{c\frac{1}{3}(2)}^2$ | .1087 | 10 | .1261 | 14 |
| $R_{cm}^2$ | .0978 | 4 | .1081 | 10 |
| (15) | .1142 | 11 | .1048 | 7 |
| (17) | .1760 | 15 | .1041 | 6 |
| (18) | .1508 | 14 | .1064 | 8 |
| (19) | .0942 | 2 | .1032 | 2.5 |
| (5&23) | .1035 | 9 | .1034 | 4 |
| (6&23) | .0982 | 5 | .1030 | 1 |
| (7&23) | .0922 | 1 | .1032 | 2.5 |
| (9&23) | .0983 | 6 | .1091 | 11.5 |
| (13&23) | .0972 | 3 | .1091 | 11.5 |
| (20&23) | .1031 | 8 | .1103 | 13 |
| (21&23) | .1011 | 7 | .1035 | 5 |

*Estimates of $\rho^2$ are ranked separately for each of the situations.

viable alternative for the researcher, $R_{c4}^2$ or $R_{cm}^2$ would be reasonable choices in terms of accuracy, but they were more variable than (18). The only conservative choice available, that is, an estimate which underestimated $\rho_c^2$, is $R_{cm}^2$.

Formula estimates and Claudy's double cross-validation procedure were compared for both multiple and stepwise regression to determine their relative accuracy and usefulness in estimating $\rho^2$. Table 2 shows the mean differences of $\rho^2$ minus the estimates for the 1978 and 1979 combined data for both multiple and stepwise regression. Each tabled value for multiple regression is the mean of approximately 120 differences and for stepwise regression is the mean of approximately 58 differences. For multiple regression, formula (9) was the most accurate, but it was also nearly the most variable. Although different by only .007 in accuracy, formula (7) was least variable. Both (9) and (7) tended to underestimate $\rho^2$ so both would be considered conservative estimates.

For stepwise regression, three of the formulas were within .0006 for standard deviations and .02 for accuracy. They were (5), (7), and (21) and they all tended to overestimate $\rho^2$. Ironically, more accurate estimates of $\rho^2$ which also underestimated $\rho^2$ were among the formulas designed to estimate $\rho_c^2$. Formulas (19), (5 & 23), and (7 & 23) had mean differences in Table 2 of .002, .002, and .003, respectively, and had standard deviations smaller than any of the $\rho^2$ estimates.

## Summary

This study was designed to compare the accuracy and variability of four cross-validation procedures and eighteen formulas in terms of their ability to estimate $\rho^2$ and $\rho_c^2$. The results of the study provide information which practical researchers, particularly those using small sample sizes, may use to determine which estimates would probably be most accurate in their research studies. Recommendations were made regarding the selection of estimates of

$\rho^2$ and $\rho_c^2$ for both multiple and stepwise regression. The researcher's choice depends upon several factors including the population value to be estimated, the type of regression used, the concern about the balance between accuracy and variability, whether cross-validation is a viable alternative in terms of time and cost, and, whether the researcher prefers a conservative approach. Since these five factors must be included in each selection of an estimate, a single estimate cannot be isolated as the recommended one for all situations.

Further research in the area should consider violations of assumptions, other types of regression procedures, control of population values to increase the range of investigation, and various methods of summarizing the data.

## APPENDIX A

1. $\hat{\rho}^2 = 1 - \frac{N(1-R^2)}{N-n}$ (Larson 1931)* (5)

2. $\hat{\rho}^2 = 1 - \frac{(N-1)(1-R^2)}{N-n}$ (Wherry 1931)* (6)

3. $\hat{\rho}^2 = 1 - \frac{(N-1)(1-R^2)}{N-n-1}$ (7)

4. $\hat{\rho}^2 = 1 - \frac{(N-3)(1-R^2)}{N-n-1}\ \frac{N-3}{N-n-1}\ \frac{2}{N-n+1}\ (1-R^2)^2$ (Herzberg 1969)* (9)

5. $\hat{\rho}^2 = 1 - \frac{(N-3)(1-R^2)}{N-n-1}\{1 + \frac{2(1-R^2)}{N-n-2.3}\}$ (Claudy 1978) (13)

6. $\hat{\rho}^2 = 1 - \frac{(N-4)(1-R^2)}{N-n-1}\{1 + \frac{2(1-R^2)}{N-n+1}\}$ (Claudy 1978) (20)

7. $\hat{\rho}^2 = R^2 - \frac{(1-R^2)}{N}\{n + 1 - (1-R^2)(1+2R^2)\}$ (Barten 1962) (21)

---

8. $\hat{\rho}_c^{\,2} = 1 - \frac{N-1}{N-n-1}\ \frac{N+n+1}{N}\ (1-R^2)$ (Herzberg 1969) (15)

9. $\hat{\rho}_c^{\,2} = 1 - \frac{N-1}{N-n-1}\ \frac{N-2}{N-n-2}\ \frac{N+1}{N}\ (1-R^2)$ (Darlington 1968)* (17)

10. $\hat{\rho}_c^{\,2} = 1 - \frac{N+n+1}{N-n-1}\ (1-R^2)$ (Uhl and Eisenberg 1970) (18)

11. $\hat{\rho}_c^{\,2} = \{\frac{NR^2-n}{R(N-n)}\}^2$ (Burket 1964)* (19)

$\hat{\rho}_c^{\,2} = \{R-2(R-\hat{\rho})\}^2 = \{2\hat{\rho}-R\}^2$ (Claudy 1978)* (23)

12. (5) and (23)
13. (6) and (23)
14. (7) and (23)
15. (9) and (23)
16. (13) and (23)
17. (20) and (23)
18. (21) and (23)

* original reference

The American College Testing Program. Highlights of the ACT technical report. Iowa City, Iowa: Author, 1973.

Barten, A. P. Note on the unbiased estimation of the squared multiple correlation coefficient. Statistica Neerlandica, 1962, 16, 151-163.

Burket, G. R. A study of reduced rank models for multiple prediction. Psychometric Monograph, 1964, 12.

Carlson, J. E. On the use of "stepwise regression." Working paper No. 15 of the Department of Educational Research, University of Pittsburgh, 1972.

Carter, D. S. Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 1979, 39, 261-266.

Claudy, J. G. Multiple regression and validity estimation in one sample. Applied Psychological Measurement, 1978, 2, 595-607.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Ezekial, M. S. B. Paper presented at the American Mathematical Society meeting. Chicago, Illinois, 1928. Cited by Wherry, 1931.

Harris, R. J. A primer of multivariate statistics. Academic Press, 1975.

Herzberg, P. A. The parameters of cross-validation. Psychometric Monograph, 1969, 16.

Kleinbaum, D. G. and Kupper, L. L. Applied regression analysis and other multivariate methods. North Scituate, MA: Duxbury Press, 1978.

Larson, S. C. The shrinkage of the coefficient of multiple correlation. Journal of Educational Psychology, 1931, 22, 45-55.

Lord, F. M. Efficiency of prediction when a regression equation from one sample is used in a new sample. Research Bulletin 50-40, Educational Testing Service, Princeton, NJ, 1950.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Co., 1968.

Mosier, C. I. Symposium. The need and means of cross-validation. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11, 12-15.

McNemar, Q. Psychological statistics (3rd ed.). New York: Wiley & Sons, 1962.

Newman, I., McNeil, K., Garver, T., and Seymour, G. A Monte Carlo evaluation of estimated parameters of five shrinkage estimate formuli. Paper presented at the Annual Meeting of AERA, April 1979, San Francisco. ED174675.

Nicholson, G. E. Jr. The application of a regression equation to a new sample. Unpublished doctoral dissertation, University of N. Carolina, 1948. (Condensed in G. E. Nicholson, Jr., Prediction in future samples. In I. Olkin et al. (Eds.) Contributions to probability and statistics. Stanford: Stanford University Press, 1960).

Olkin, I. and Pratt, J. Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 1958, 29, 201-211.

Pratt, J. W. Personal communication to E. E. Cureton, October 20, 1964. Cited in Claudy, 1978.

Pratt, J. W. Personal communication to J. G. Claudy, May 12, 1969. Cited in Claudy, 1978.

Randall, W. Personal communication, June 1980.

Schmitt, N., Coyle, B., and Rauchenberg, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 1977, 84, 751-758.

Stein, C. Multiple Regression. In I. Olkin et al. (Eds.) Contributions to probability and statistics. Stanford: Stanford University Press, 1960.

Uhl, N. and Eisenberg, T. Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 1970, 30, 487-489.

Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 1931, 2, 440-457.

Wherry, R. J. Appendix V. In W. H. Stead and C. P. Shartle (Eds.) Occupational counseling techniques. New York: American Book Company, 1940, 245-252.

Cummings, C.C. Cross-validation and formula estimates of the shrinkage of the multiple correlation coefficient. Unpublished doctoral dissertation, University of Pittsburgh, 1980.